Exponential Capacity in an Autoencoder Neural Network with a Hidden Layer

Alireza Alemi*, Alia Abbara
Limits of computations in deep nets?

random binary input patterns

random binary output labels
Limits of computations in deep nets?

Capacity = \frac{\text{# input-output associations}}{\text{network size}}

random binary input patterns

input layer

hidden layer 1

hidden layer 2

hidden layer 3

output layer

random binary output labels

Alireza Alemi  PIML 2018
Capacity of multi-layer perceptron (MLP) with one hidden layer?
Investigating internal representation

- Input layer ($N_v$ neurons)
- Hidden layer ($N_h$ neurons)
- Output layer

Patterns $\{\xi^\mu\}$

Random binary output labels
Expansive (or overcomplete) representation

\[ \Lambda = \frac{N_h}{N_v} > 1 \]
Expansive (or overcomplete) representation

<table>
<thead>
<tr>
<th>Nervous system</th>
<th>Expansion ratio ($\Lambda$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Entorhinal cortex $\rightarrow$ Dentate gyrus</td>
<td>$\sim 10$</td>
</tr>
<tr>
<td>LGN $\rightarrow$ V1</td>
<td>$\sim 25$</td>
</tr>
<tr>
<td>in fly olfactory system:</td>
<td></td>
</tr>
<tr>
<td>antennal lobe $\rightarrow$ mushroom body</td>
<td>$\sim 40$</td>
</tr>
<tr>
<td>in the cerebellum, mossy fibers $\rightarrow$ granule cells</td>
<td>$100 \sim 200$</td>
</tr>
<tr>
<td>rodent’s olfactory bulb $\rightarrow$ piriform cortex</td>
<td>$\sim 10^3$</td>
</tr>
</tbody>
</table>
Expansive representation with random weights

input layer ($N_v$ neurons)  hidden layer ($N_h$ neurons)  output layer

patterns $\{\xi^\mu\}$

random fixed weights $v_{ij}$

random binary output labels
Expansive representation with random weights

input layer
($N_v$ neurons)

hidden layer
($N_h$ neurons)

output layer

patterns
{$\xi^\mu$}

where $\mu=1\ldots p$

random fixed weights $v_{ij}$

random binary output labels

How good is random-weight expansive representations?
Expansive representation with random weights

- Input layer ($N_v$ neurons)
- Hidden layer ($N_h$ neurons)
- Output layer

Patterns $\{\xi^\mu\}$ where $\mu = 1 \ldots p$

How many of input patterns can be reconstructed from this representation?
Expansive autoencoder with random weights

Focus of this talk: Optimal capacity of expansive autoencoder?
Critical capacity of the perceptron: Gardner approach

\[ y^\mu = \text{sgn}\left( \frac{\mathbf{w} \cdot \xi^\mu}{\sqrt{N}} - \theta \right) \]

- Find \( \mathbf{w} \) to store a set of random ensemble of input-output pairs \( \{ (\xi^\mu, d^\mu) \} \), \( \mu=1 \ldots p \) with a robustness \( \kappa \)

- Storage criteria: \( \forall \mu, i : d^\mu \left( \frac{1}{\sqrt{N}} \sum_{i=1}^{N} w_i \xi_i^\mu - \theta \right) > \kappa \)

Gardner volume (\( N \rightarrow \infty \)):

\[ \Omega = \int_{\|\mathbf{w}\|^2=N} d^N \mathbf{w} \prod_{\mu=1}^{p} \Theta\left( d^\mu \left( \frac{1}{\sqrt{N}} \sum_{i=1}^{N} w_i \xi_i^\mu - \theta \right) - \kappa \right) \]

\( \llangle \log(\Omega) \rrangle \) averaging over the (quenched) distribution of the patterns

\[ \llangle \log(\Omega) \rrangle = \lim_{n \to 0} \frac{\llangle \Omega^n \rrangle - 1}{n}, \text{ critical capacity: } \alpha_c = \lim_{N \to \infty} \frac{p_c}{N} \]

- Perceptron learning rule (\( \kappa=0 \)): \( \Delta w_i = \eta \xi_i (d^\mu - y^\mu) \)
A simplified expansive (over-complete) autoencoder

- Binary neurons ($\pm 1$)
- Expansion ratio: $\Lambda = N_h / N_v$
- Fixed encoding weights $v_{ij}$ are randomly sampled from $\mathcal{N}(0, 1)$
- Goal: reconstruct random patterns $\{\xi^\mu\}_{\mu=1,...,p}$ by learning the decoding weights $w_{ji}$
- Perceptron learning rule: $\Delta w_{ji} = \eta(\xi_j^\mu - y_j^\mu)\sigma_i^\mu$
- Dense coding level for patterns: $P(\xi_i=+1) = 0.5$
- Coding level of hidden layer: $P(\sigma_i=+1) = f$
- Maximal/critical storage Capacity: $\alpha_{\text{max}} = p_{\text{max}} / N_h$ as $N_h \rightarrow \infty$
- Storage criteria: meeting fixed-point equations

$$\forall j, \mu: \xi_j^\mu = \text{sgn} \left( \sum_i w_{ji} \text{sgn} \left( \sum_l v_{il} \xi_l^\mu \right) \right)$$

Dynamics:

$$\sigma_i^\mu = \text{sgn} \left( \sum_{j=1}^{N_v} v_{ij} \xi_j^\mu - \theta \right)$$

$$y_j^\mu = \text{sgn} \left( \sum_{i=1}^{N_h} w_{ji} \sigma_i^\mu \right)$$

Alemi A., Abbara A., 2017 ArXiv
A mean-field approximation (MFA)

\[ \xi_j^\mu = \text{sgn} \left( \sum_{i=1}^{N_h} w_{ji} \text{sgn} \left( \sum_{l=1}^{N_v} v_{il} \xi_l^\mu \right) \right) \]

\[ = \text{sgn} \left( \sum_{i=1}^{N_h} w_{ji} \text{sgn} \left( \sum_{l \neq j; l=1}^{N_v} v_{il} \xi_l^\mu + v_{ij} \xi_j^\mu \right) \right) \]

\[ = \text{sgn} \left( \sum_{i=1}^{N_h} w_{ji} \text{sgn} \left( z_i^\mu + v_{ij} \xi_j^\mu \right) \right) \]

\[ \text{Approximated by a } quenched \text{ random variable with a Gaussian distribution: } \mathcal{N}(0, N_v) \]

\[ P(\sigma_i^\mu | \xi_j^\mu) \approx \frac{1}{2} + \sigma_i^\mu \xi_j^\mu \frac{v_{ij}}{\sqrt{2\pi N_v}} \]

\[ (\sigma_i^\mu \perp \sigma_k^\mu) | \xi_j^\mu \implies P(\sigma^\mu | \xi_j^\mu) = \prod_i P(\sigma_i^\mu | \xi_j^\mu) \]
A mean-field approximation (MFA)

\[ P(\sigma_i^\mu | \xi_j^\mu) \approx \frac{1}{2} + \sigma_i^\mu \xi_j^\mu \frac{v_{ij}}{\sqrt{2\pi N_v}} \]

\[ (\sigma_i^\mu \perp \sigma_k^\mu) \mid \xi_j^\mu \implies P(\sigma^\mu | \xi_j^\mu) = \prod_i P(\sigma_i^\mu | \xi_j^\mu) \]
Mean-Field Approximation (MFA) for the \textit{replica} calculation

Gardner volume:

\[ \Omega = \int_{\| \mathbf{w} \|^2 = N_h} d^{N_h} \mathbf{w} \prod_{\mu=1}^{p} \Theta(\xi_j^\mu - \kappa) \]

Replica theory: compute \( \langle \log(\Omega) \rangle \) over distribution of patterns using the replica method

\[ \langle \log \Omega \rangle = \lim_{n \to 0} \frac{\langle \Omega^n \rangle - 1}{n} \]

Probability distribution

\[ P(\sigma_i^\mu | \xi_j^\mu) = \frac{1}{2} + \sigma_i^\mu \xi_j^\mu \frac{v_{ij}}{\sqrt{2\pi N_v}} \]

\[ (\sigma_i^\mu \perp \sigma_k^\mu) | \xi_j^\mu \implies P(\sigma^\mu | \xi_j^\mu) = \prod_i P(\sigma_i^\mu | \xi_j^\mu) \]
Exponential critical capacity in the MFA

Solutions (replica-symmetric, locally stable):

\[ \alpha_{c \text{MFA}} \left( \int_{M \sqrt{2\Lambda/\pi} - \kappa}^{\infty} \right) D_t \left( \kappa + t - \sqrt{2\Lambda/\pi} \right)^2 = 1 - M^2 \]

\[ \alpha_{c \text{MFA}} \left( \int_{M \sqrt{2\Lambda/\pi} - \kappa}^{\infty} \right) D_t \left( \kappa + t - \sqrt{2\Lambda/\pi} \right) \sqrt{2\Lambda/\pi} = M, \]

where

\[ M = \sum_i \frac{v_i w_i}{N_h}, \quad \Lambda = \frac{N_h}{N_v} \]

\[ D_t \equiv \frac{dt}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} \]

Analytical mean-field

Exponential fit
Comparison with simulation (perceptron learning rule)

![Graph showing comparison between simulation and analytical mean-field for capacity and symmetry between $w_{ji}$ and $v_{ij}$ as a function of expansion ratio ($\Lambda$).](image)

Expansion ratio ($\Lambda$)

Capacity ($\alpha$)

Simulation mean-field
Analytical mean-field

Symmetry between $w_{ji}$ & $v_{ij}$

Simulation
Analytical

Alemi A., Abbara A, 2017 ArXiv
Effect of robustness of output, sparseness in the hidden layer

Alemi A., Abbara A, 2017 ArXiv
Comparison with the full-model

![Graph showing comparison between expansion ratio and capacity for full model and mean-field simulation.

- Simulation full model
- Simulation mean-field

Capacity (\(\alpha\)) vs Expansion ratio (\(\Lambda\))

- Log-log scale

Alemi A., Abbara A, 2017 ArXiv
Log of number of patterns vs. total number of neurons
Discussion

• Expansive representation has important computational implications.
• *Exponential* scaling of the capacity with the expansion ratio in our autoencoder
  – Expansion makes up for the loss of information due to binarization
• In very good agreement with results of simulations using an online learning rule
• Future directions
  – Robustness to noise by training encoding weights
  – Computing the capacity of the full autoencoder model
  – Extension to the capacity of MLP and deep nets
Acknowledgement

Alia Abbara

Discussion with many colleagues ...
Question?